Hypothesis Testing

Shannon Pileggi

STAT 217

OUTLINE

Overview

Statistical inference is drawing conclusions about a population based on a sample from that population. Two methods of statistical inference are

- *confidence intervals* - estimate a population parameter (with a range of values)
- *hypothesis testing* - assess evidence for a specific (single) value for a population parameter

The population parameters we are going to analyze include:

- one mean
- mean of a difference
- two means
- more than two means
- one proportion
- N/A for STAT 217
- two proportions
- more than two proportions

Hypothesis testing

A **hypothesis test** is a method of using data to **summarize the evidence about a hypothesis**.

My 7 steps for significance test about a hypothesis:

1. Define the parameter of interest
2. Hypotheses
3. Identify the appropriate test
4. Assumptions
5. Test Statistic
6. *p*-value
7. Conclusion

H_0 and H_a

- We start with a null hypothesis (H_0) that represents the status quo. This is generally a statement of no effect.
- The research question is represented in the alternative hypothesis (H_a). This is a statement of effect, and it is what we are hoping to show.
- We conduct the test assuming the null hypothesis is true.
- We then evaluate the test results to determine if there is enough evidence to reject the null in favor of the alternative (what we hope to show).
- Translating a research question into null and alternative hypotheses can be tricky.

1. Define the parameter of interest

So far, we are focusing on two types of parameters:

- μ = population mean
- p = population proportion

in the *context* of the research question at hand.

- Research question: Data shows that UCSB students spend \$52 on average for a haircut. Do Cal Poly students spend the same?
- Parameter: μ = population mean amount spent on a haircut by Cal Poly students

- Research question: A professor claims that 70% of Cal Poly students own a Mac laptop. Can we find evidence for or against this?
- Parameter: p = population proportion of Cal Poly students that own a Mac laptop

Example research question

A professor claims that 70% of Cal Poly students own a Mac laptop. When we examine our sample data, we see that $\hat{p} = 50/67 = 0.75$, or 75% of students own a Mac laptop. Do we have evidence for or against this claim?

Two possible explanations

- The population proportion really is different than 0.7
- The population proportion is actually 0.7, but we observed 0.75 by random chance.

This is why we do hypothesis testing!

2. State the hypotheses

A professor claims that 70% of Cal Poly students own a Mac laptop. When we examine our sample data, we see that $\hat{p} = 50/67 = 0.75$, or 75% of students own a Mac laptop. Do we have evidence for or against this claim?

- The null hypothesis is one of no effect - that the population proportion of Cal Poly students that own a Mac laptop is 0.7.

  H_0: $p = 0.7$
- The alternative hypothesis is one of an effect - that the population proportion of Cal Poly student that own a Mac laptop is different than 0.7.

  H_a: $p \neq 0.7$

Forms of the hypotheses

Hypotheses are always about a population parameter.

Null Hypothesis

I think that the population proportion of Cal Poly students that own a Mac laptop is 0.7. H_0: $p = 0.7$

Alternative Hypothesis

I think that the population proportion of Cal Poly students that own a Mac laptop differs from 0.7. H_a: $p \neq 0.7$ two-sided

I think that the population proportion of Cal Poly students that own a Mac laptop is less than 0.7. H_a: $p < 0.7$ one-sided

I think that the population proportion of Cal Poly students that own a Mac laptop is greater than 0.7. H_a: $p > 0.7$ one-sided

Group Exercise

Planned Parenthood wanted to see if a majority of Cal Poly students would support allowing those under the age of 18 to have access to birth control pills without their parents permission.

What would be the null and alternative hypothesis of their study?

1. H_0: $p = 0$ vs H_a: $p \neq 0$
2. H_0: $p = 0.5$ vs H_a: $p > 0.5$
3. H_0: $p > 0.5$ vs H_a: $p = 0.5$
4. H_0: $\mu = 0.5$ vs H_a: $\mu > 0.5$
5. H_0: $\mu > 0.5$ vs H_a: $\mu = 0.5$

Group Exercise

200 Cal Poly students were asked how many colleges they applied to: the sample had an average of 9.7 college applications with a standard deviation of 7. The College Board website states that counselors recommend students apply to roughly 8 colleges.

Which of the following are the correct set of hypotheses to test if these data provide convincing evidence that the average number of colleges all Cal Poly students apply to is higher than recommended?

1. H_0: $\mu = 9.7$ vs H_a: $\mu > 9.7$
2. H_0: $\mu = 8$ vs H_a: $\mu > 8$
3. H_0: $\bar{x} = 8$ vs H_a: $\bar{x} > 8$
4. H_0: $\mu = 8$ vs H_a: $\mu > 9.7$
5. H_0: $\mu > 8$ vs H_a: $\mu = 9.7$

Group Exercise

In 2000, research indicated that the proportion of Americans who were vegetarian was 0.15. A recent random sample of 1500 Americans shows that the proportion of Americans who are vegetarian is 0.17. We would like to know the proportion of Americans who are vegetarian has changed since 2000.

What would be an appropriate null and alternative hypothesis?

1. H_0: $\hat{p} = 0.15$ vs H_a: $\hat{p} \neq 0.15$
2. H_0: $p = 0.15$ vs H_a: $p \neq 0.15$
3. H_0: $\hat{p} = 0.17$ vs H_a: $\hat{p} \neq 0.17$
4. H_0: $p = 0.17$ vs H_a: $p \neq 0.17$
5. H_0: $\bar{x} = 0.17$ vs H_a: $\bar{x} \neq 0.17$
6. H_0: $\mu = 0.15$ vs H_a: $\mu \neq 0.15$

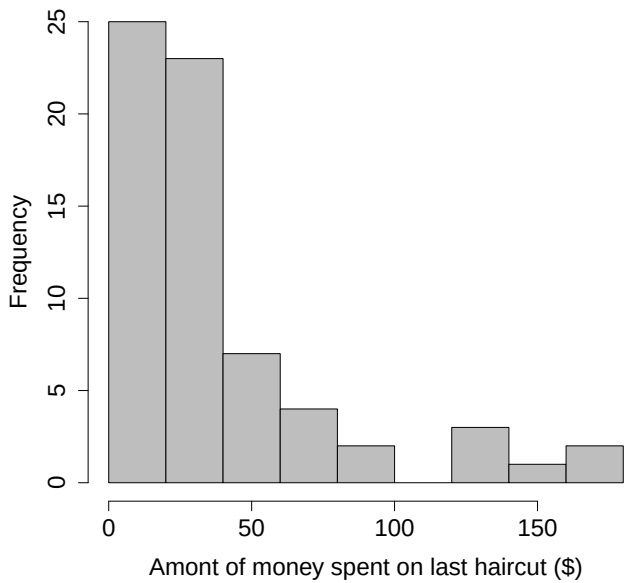
1. Parameter of interest

- Research question: Data shows that UCSB students spend \$52 on average for a haircut. Do Cal Poly students less than that?
- Parameter: μ = true population mean amount spent on a haircut by Cal Poly students

Data shows that UCSB students spend \$52 on average for a haircut. Do Cal Poly students less than that?

What is an appropriate null hypothesis for this test?

1. $H_0 : \mu = 52$
2. $H_0 : \bar{x} = 40$
3. $H_0 : \mu = 40$
4. $H_0 : \bar{x} = 52$
5. $H_0 : p = 52$
6. $H_0 : p = 40$
7. $H_0 : \mu \neq 52$
8. $H_0 : \mu < 52$



$n = 50,\ \bar{x} = 40,\ s = 25$

2. Hypotheses

parameter of interest: μ

- The null and alternative hypotheses are:

  H_0: $\mu = 52$ vs H_a: $\mu < 52$
- The value of μ in our hypothesis statement is called μ_0. Here, $\mu_0 = 52$. This is what we assume to be true about μ **under** the null hypothesis.

3. Identify the appropriate test

What is the appropriate statistical test?

- hypothesis test about a mean
- one-sample t-test
- t-test

These are all the same test by different names.

4. Conditions

Are conditions met for this test?

- The observations are independent.
- The population distribution is normal *or* we have a 'large' sample size ($n \geq 30$).

(These are the same conditions required for a CI for μ.)

H_0: $\mu = 52$ vs H_a: $\mu < 52$

Which of the following sample means do you think would provide the most evidence against H_0?

1. $\bar{x} = 20$
2. $\bar{x} = 30$
3. $\bar{x} = 40$
4. $\bar{x} = 50$
5. not enough information to determine

5. The test statistic

The test statistic looks like a z-score: $z = \dfrac{\text{value} - \text{mean}}{\text{standard deviation}}$

$$\text{test statistic} = \frac{\text{sample mean} - \text{null hypothesis mean}}{\text{standard error of the sample mean}}$$

$$t = \frac{\bar{x} - \mu_0}{se_{\bar{x}}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

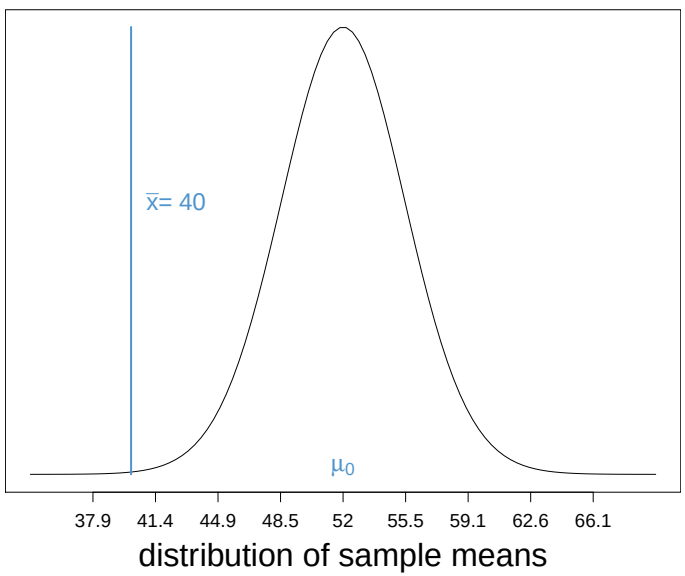
This is a **t test statistic** because the test statistic follows a t distribution with $n - 1$ degrees of freedom.

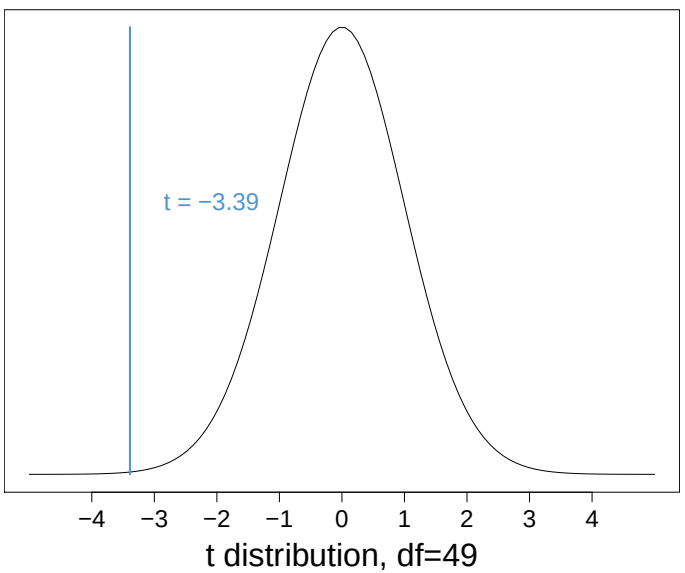
5. The test statistic

H_0: $\mu = 52$ vs H_a: $\mu < 52$, $n = 50$, $\bar{x} = 40$, $s = 25$

$t =$

This particular t test statistic follows a t distribution with 49 degrees of freedom.





5. The test statistic

A **test statistic** describes how far our observed sample statistic falls from the parameter value given in the null hypothesis. How? It is the number of standard errors between the observed sample statistic and the parameter value given in the null hypothesis.

If the test statistic falls 'far' from the value suggested by the null hypothesis, it is evidence against the null hypothesis and in favor of the alternative hypothesis.

H_0: $\mu = 52$ vs H_a: $\mu < 52$

Which of the following test statistics do you think would provide the most evidence against H_0?

1. $t = -1$
2. $t = -2$
3. $t = -3$
4. $t = -4$
5. not enough information to determine

6. p-value

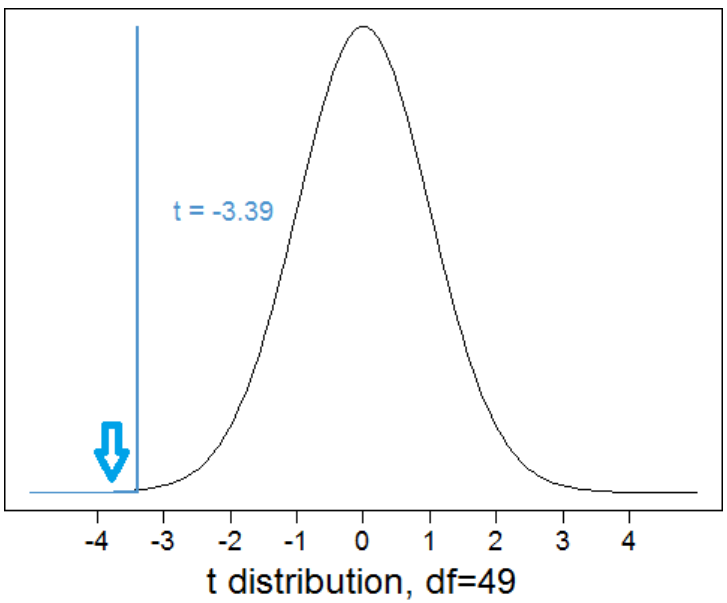
- We *quantify* the evidence against the null hypothesis in favor of the alternative with a p-value.
- The p-value is the **probability** that the test statistic takes the observed value or a value more extreme, *assuming H_0 is true*.
- Since the p-value is always calculated under the assumption that H_0 is true, we always interpret this as evidence either supporting or not supporting the null hypothesis.
- **The smaller the p-value the stronger the evidence against H_0.**

6. The p-value

The p-value is a one-tail probability from the t-distribution because we had a one-sided H_a. This is given by the area under the curve of a t distribution with 49 degrees of freedom less than the value of the test statistic (because H_a was *less* than).

$p\text{-value} = Pr(t_{49} < -3.39)$

In R, we can calculate this p-value exactly: $p\text{-value} = 0.0006$



Results in R

```
>t.test(survey$Haircut,mu=52,alternative="less")

        One Sample t-test

data:  survey$Haircut
t = -3.3941, df = 49, p-value = 0.0006
alternative hypothesis: true mean is less than 52
95 percent confidence interval:
 32.8935 47.1064
sample estimates:
mean of x
  40.0123
```

6. p-value

- All p-values are probabilities.
- The smaller the p-value the stronger the evidence against H_0.
- A small p-value means that the sample data would be unusual if H_0 were true.
- When the p-value is 'small' you may conclude that there is strong evidence against H_0.
- When the p-value is not 'small' you may conclude that there is not strong evidence against H_0. This implies that the null hypothesis is plausible.

Elements of the *p*-value interpretation

- To interpret the *p*-value, apply the definition of the *p*-value to the context of your research question.
- *Definition:* the *p*-value is the probability that the test statistic takes the observed value or a value more extreme, assuming H_0 is true..
- *Interpretation:* If the average amount spent on a hair cut by Cal Poly students was truly equal to 52, the probability that we would observe test statistic as extreme or more extreme than -3.39 is 0.0006.

More ways we can think about the *p*-value

- If the average amount spent on a hair cut by Cal Poly students was truly equal to 52, the probability that we would observe test statistic as extreme or more extreme than -3.39 is 0.0006.
- If the average amount spent on a hair cut by Cal Poly students was truly equal to 52, the probability that we would observe a sample mean less than or equal to 40 is 0.0006.
- That is, there is about a 0.06% chance we would observe a test statistic that extreme or more extreme if the null hypothesis were true.
- If we were to take 10,000 random samples from a population where the average amount spent on a hair cut was truly equal to 52, only 6 out of those 10,000 samples would have results as extreme as the one we observed in our sample.
- This **is** substantial evidence *against* H_0; therefore, we have evidence in favor of H_A.

H_0: $\mu = 52$ vs H_a: $\mu < 52$

Which of the following p-values do you think would provide the most evidence against H_0?

1. $p = 0.02$
2. $p = 0.23$
3. $p = 0.56$
4. $p = 0.72$
5. not enough information to determine

Making a decision

Based on the *p*-value, make a decision about H_0 based on a pre-specified **level of significance**, also known as α. We always select α *before* looking at the data, and α is commonly set at 0.05.

Possible decisions:	
1. reject H_0,	*p*-value $\leq \alpha$
2. fail to reject H_0,	*p*-value $> \alpha$

Draw a number line:

For which of the following p-values would you reject H_0 at the $\alpha = 0.05$ level of significance?

1. $p = 0.02$
2. $p = 0.23$
3. $p = 0.56$
4. $p = 0.72$
5. not enough information to determine

Starting a conclusion

p-value $\leq \alpha$

- *decision:* reject H_0
- *conclusion:* we **do** have evidence in favor of H_a
- *example:* We **do** have evidence that the population mean amount spent on a haircut is less than \$52.

p-value $> \alpha$

- *decision:* fail to reject H_0
- *conclusion:* we **do not** have evidence in favor of H_a
- *example:* We **do not** have evidence that the population mean amount spent on a haircut is less than \$52.

Statically significant

- When an 'effect' is so large that it would be rare to see such a difference by ordinary random variation, we say that the results are *statistically significant*.
- When p-value $\leq \alpha$ and we reject H_0, we say that the result is *statistically significant*.
- When p-value $> \alpha$ and we fail to reject H_0, we say that the result is *not statistically significant*.

Result	**Interpretation**	**Decision**	**Stat. Sig.?**
p-value $\leq \alpha$	evidence in favor of H_a	reject H_0	yes
p-value $> \alpha$	no evidence in favor of H_a	fail to reject H_0	no

7. Conclusion

Elements of a conclusion:

1. **Decision about H_0:**

   At the 0.05 level of significance, we reject the null hypotheses that the average amount spent on a haircut is equal to \$52.

2. **Statement about the parameter tested in context of the research question:**

   We do have evidence that the population mean amount of money Cal Poly students spend on a hair cut is less than \$52.

3. **Provide a deeper connection of how this relates to the research question:**

   Because we have evidence that the population average amount spent on a haircut by Cal Poly students is less than \$52, this means that Cal Poly students tend to spend less than UCSB students, on average.

Going further...

We found evidence that the population mean amount of money Cal Poly students spend on a hair cut is less than \$52. How much less? Is it meaningfully less than \$52?

The form of your alternative hypothesis determines how you calculate your p-value.

You should establish your null and alternative hypotheses before you look at your data.

What happens to the value of your test statistic and p-value when different elements of your data change?

$$t = \frac{\bar{x} - \mu_0}{se_{\bar{x}}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

- $n \uparrow$ se t p
- $s \uparrow$ se t p
- $|\bar{x} - \mu_0| \uparrow$ se t p

The *p*-value is

1. the probability that the null hypothesis is true.
2. the probability that the null hypothesis is false.
3. the probability that the test statistic had the observed result or more extreme if the null hypothesis were true.
4. the probability that the test statistic had the observed result or more extreme if the alternative hypothesis were true.
5. A and C
6. B and D

Interpreting your decision for one-sided H_a

Consider testing that the average GPA of Cal Poly students is less than 3.50. (H_0: $\mu = 3.5$ vs H_a: $\mu < 3.5$)

Write out a decision tree for 'small' and 'large' p-values:

Interpreting your decision for two-sided H_a

Consider testing that the average GPA of Cal Poly students differs from 3.50. (H_0: $\mu = 3.5$ vs H_a: $\mu \neq 3.5$)

Write out a decision tree for 'small' and 'large' p-values:

Interpreting your decision: fail to reject H_0

- fail to reject $H_0 \neq$ accept H_0!
- **never** accept H_0!
- Why? Some context helps...
  Consider H_0: $\mu = 3.5$ vs H_a: $\mu \neq 3.5$

```
data:  gpa
t = 0.43674, df = 49, p-value = 0.6642
alternative hypothesis: true mean is not equal to 3.5
95 percent confidence interval:
 3.256276 3.879075
sample estimates:
mean of x
 3.567676
```

Public health officials are concerned about the water crisis in Flint, Michigan, where lead is seeping into the water supply. The CDC recommends that an acceptable level of lead in blood is 5 micrograms per deciliter. The officials take blood samples from 20 children from Flint, Michigan and find that the average lead blood level is $\bar{x} = 8.3$; they test H_0: $\mu = 5$ vs H_a: $\mu \neq 5$ and get a p-value of 0.034. What can be said at the $\alpha = 0.05$ level of significance?

We (do/do not) have statistically significant evidence that the (population/sample) mean is (different from/equal to/greater than) 5 micrograms per deciliter.

1. do; population; different than
2. do; sample; different than
3. do; population; greater than
4. do not; sample; greater than
5. do not; population; equal to

Study 25-35

Suppose I want to determine if students follow Dean Bailey's guidelines to study 25-35 hours a week. I take data from a sample of students, and test the hypotheses H_0: $\mu = 30$ vs H_a: $\mu \neq 30$. I get a p-value $= 0.56$ so I fail to reject H_0.

A confidence interval for μ would...

1. contain 30
2. not contain 30
3. contain 0.56
4. not contain 0.56
5. not enough information to determine

Agreement of confidence intervals and hypothesis tests

Confidence intervals and hypothesis tests (almost) always agree, as long as the two methods use *equivalent* levels of significance / confidence. A **two-sided** hypothesis with α level of significance is equivalent to a $100 \times (1 - \alpha)\%$ CI. Agreement examples:

α	confidence level
0.01	
0.05	95%
0.10	

- When you reject H_0 the corresponding CI **should not** include value in H_0.
- When you fail to reject H_0 the corresponding CI **should** include value in H_0.

Group Exercise

Suppose I performed a hypothesis test at the $\alpha = 0.01$ level of significance about the average GPA of Cal Poly students as follows: H_0: $\mu = 3.55$ vs H_a: $\mu \neq 3.55$. The test yielded a p-value of 0.03.

Using the same data, what can we say about a 99% confidence interval for μ =true mean GPA of Cal Poly students?

1. a 99% CI for μ would contain 3.55
2. a 99% CI for μ would not contain 3.55
3. a 99% CI for μ would contain 0.03
4. a 99% CI for μ would not contain 0.03
5. there is not enough information to determine